# TEST PROCEDURES

# FOR THE

# MARCH 1987 DARPA BENCHMARK TESTS

David S. Pallett

Institute for Computer Sciences and Technology
National Bureau of Standards
Gaithersburg, MD 20899

## ABSTRACT

This paper describes test procedures that were to be used in conducting benchmark performance tests prior to the March 1987 DARPA Meeting. These tests were to be conducted using selected speech database material and input from "live talkers", as described in a companion paper.

## INTRODUCTION

At the Fall 1986 DARPA Speech Recognition Meeting, plans were discussed for implementing benchmark tests using the Task Domain Speech Database. There was additional discussion of the desirability of developing and implementing "live tests" using speech material provided by speakers at the contractors' facilities, emulating in some sense the process of inputting speech material during a demonstration of real-time performance. Following the Fall Meeting, the Task Domain Speech Database was recorded at TI and significant portions of it were made available for system development and training purposes through NBS to both CMU and BBN. Another portion was selected for use in implementing these benchmark tests [1], and this test material was distributed to CMU and BBN during the last week of February, 1987. This paper outlines test procedures to be used to implement these tests prior to the March 1987 Meeting.

A number of informal documents have circulated within the DARPA Speech Recognition community that outline proposed test procedures. A Strategic Computing draft document dated Dec. 6, 1985 [2] identified key issues in some detail. Portions of this were heavily annotated and distributed to several sites during June 1986 and were the subject of discussions involving the author and representatives of CMU, BBN, Dragon Systems, MIT and TI during visits during June and early July 1986. These discussions were valuable in developing an outline of benchmark test procedures [3] that was discussed at the Fall 1986 DARPA Meeting, and which was structured after a model for performance assessment tests outlined in an earlier NBS publication [4]. Thus the present proposed test procedure represents the most recent and specifically focussed in a series of documents outlining test procedures for the DARPA Speech Recognition Program.

## EXPERIMENTAL DESIGN

There were to be two distinct types of tests conducted prior to the March 1987 DARPA meeting:

(1) Tests based on use of a subset of the Task Domain (Resource Management) Development Test Set Speech Database. This subset was to include use of 100 sentence utterances in either the Speaker Independent or Speaker Dependent portions of the database. The process of selecting speakers and the specific utterances is described in Reference [1]. In each case, there was considerable freedom to choose system-dependent factors such as the amount of training material for Speaker Dependent technology and the most appropriate grammar. All of the 100 specified test sentences were to be processed and reported on at the meeting. "Spell-mode" material (spelled-out representations of the letter strings for items in the lexicon) was available for use, but it processing this material was not required.

These sentence utterances were to be processed both with and without the use of imposed grammars. In the case of using no grammar, the perplexity is essentially to be nominally 1000. Comparable detailed results are to be reported for both conditions. No other parameters are to be changed for these comparitive tests.

Optionally, the same data may be processed using the "rapid adaptation" sentences for system adaptation. There is to be no use of adaptation during processing of the test material.

(2) Tests based on input provided from "live talkers". The test talkers visited both CMU and BBN prior to the March meeting. Each of the talkers spoke the "rapid adaptation" sentences and read a script containing 30 sentences drawn from the task domain sentence corpus. Data derived from the input from "live talkers" was to be analyzed and reported on at the March meeting.

LIVE TEST PROTOCOL

The microphone was to be the same as that used at TI for the Resource Management database, the Sennheiser HMD 414-6. This is a headset-mounted noise cancelling microphone similar to the Shure SM-10 family of microphones. The headset is a supra-aural headset that allows the subject to be aware of nearby conversation or instructions for prompting. The test environment was to be a conference room or computer lab. There was to be no background speech at the time the test material is provided. Test utterances could be rejected (and the subject asked to repeat the sentence) if in the judgement of the person(s) administering the tests there was some noise artifact (e.g. coughs or paper-shuffling noises) or severe mis-articulation of the test sentence. Evidence of this could be obtained by play-back of the digitized utterance.

For systems that require time to develop speaker-adaptive models, the subjects were to provide the 10 "rapid adaptation" sentences prior to the tests (e.g. the evening prior to the tests).

For one of the speakers, the 30 test sentences were to be read in and processing (automatic recognition) could take place "off-line". For the other two speakers, the test sentences were to be read in, one at a time, waiting for the system to recognize each sentence before proceeding to the next sentence. At the end of 30 minutes, if all 30 sentences had not been read in and recognized, the remaining sentences were to be read in for "off-line" processing. In practice, only three to five sentences were recognized interactively within the 30 minute period, and the remaining sentences were then read in. The elapsed time for each speaker providing the test material in this manner was typically 45 minutes. If requested, each speaker was to read in 10 words randomly chosen from the "spellmode" vocabulary subset.

PROCESSING OF LIVE INPUT

The systems were to process the test material in a manner similar to that used for the Resource Management database test material. Statistics comparable to those for the 100 sentence subsets were to be prepared and reported on at the March meeting.

ADAPTATION

Although the use of the "rapid adaptation" sentences was to be permitted, it appears that the only use made of the rapid adaptation sentences was in adapting the Speaker Dependent system at BBN for the "live test" speakers.

There was to be no use of any of the test material to enroll, adapt or to optimize system performance for the test material through repeated analyses and re-use of the test material. Intended allowable exceptions to this prohibition against re-use of the test material include demonstrating the effects of using different grammars, different strategies for enrollment, different algorithms for auditory modelling, acoustic-phonetic feature extraction, different HMM techniques, system architectures, etc. It is recognized that the breadth of these exceptions in effect limit the future use of this test material, since such extensive use of test material to demonstrate parametric effects constitutes training on test material.

Since a finite set of task domain sentences was developed at BBN, and the entire corpus of task domain sentences was made available to both CMU and BBN, in some cases the grammars used for these tests have been adapted to this finite set of sentences, including the test material.

VOCABULARY/LEXICON/OUTPUT CONVENTIONS

The task domain sentences in effect define the vocabulary. Internal representations (lexicon entries) may be at the system designer's choice, but for the purposes of implementing uniform scoring procedures, a convention was defined, drawing on material provided by CMU [5], BBN and TI. This convention includes the following considerations:

Case differences are not preserved. All input (reference) strings and output strings are in upper case.

There is no end-of-sentence punctuation. Nor is there any required special symbol to denote silences (either pre-pended, within the sentence utterance, or appended) or to indicate failure of a system to parse the reference string or input speech.

Apostrophes are represented by plusses. Words with apostrophes (embedded or appended) are represented as single words. Thus "it's" becomes "IT+S".

Abbreviations become single words. All periods indicating abbreviations are removed and the word is closed up (e.g. "U. S. A." becomes "USA").

Hyphenated items count as single words. In general, compound words that do not normally appear as separate words in the context of the assumed task domain model are entered as single, hyphenated items. The exception to this rule are compounds that include a geographic term, such as STRAIT, SEA or GULF. Thus entries such as the following count as single "words": HONG-KONG, SAN-DIEGO, ICE-NINE, PAC-ALERT, LAT-LON, PUGET-1, M-RATING, C-CODE, SQQ-23, etc. However, BERING STRAIT is to count as two words since this compound includes the geographic term "STRAIT", and it is not to be hyphenated.

Acronyms count as single words, and the output representation is not the form of the acronym made easier to interpret or pronounce (e.g. "PACFLT", not PAC-FLEET or PAC FLEET).

Mixed strings of alpha-numerics are treated as acronyms. Thus, "A42128" is treated as a one-word acronym, even though the prompt form of this indicates that this is to be pronounced as "A-4-2-1-2-8". Strings of the alpha set are also treated as acronyms (e.g. "USA"). Strings of digits are entered in a manner that takes into account the context in which they appear. Thus for a date such as 1987, it is represented as three words: "NINETEEN" "EIGHTY" "SEVEN". If it is referred to as a cardinal number it would be represented as "ONE" "THOUSAND" "NINE" "HUNDRED" "EIGHTY" "SEVEN".

## SCORING THE TEST MATERIAL

For results to be reported at the March meeting, the use of different scoring software will be acceptable. Each contracter was free to use software consistent with the following general requirements:

Data are to be reported at two levels: sentence level and word level.

At the sentence level, a sentence is to be reported as correctly recognized only if all words are correctly recognized and there are no deletion or insertion errors (other than insertions of a word or symbol for silence or a pause). The percent of sentences correctly recognized is to be reported, along with the percent of sentences that contain (at least one) insertion error(s), the percent of sentences that contain (at least one) deletion error(s) and the percent of sentences that contain (at least one) substitution error(s). The number to be used for the denominator in computing these percentages is the number of input sentences in the relevant test subset, without allowing for rejection of sentences or utterances that may not parse or for which poor scores result.

At the word level, data that are to be reported include the percent of words in the reference string that have been correctly recognized. For these tests, "correct recognition" does not require that any criterion be satisfied with regard to word beginning or ending times. It is valuable, but not required, to report the percent of insertion, deletion, and substitution errors occurring in the system output.

For those systems that provide sentence or word lattice output, scoring should be based on the top-ranked sentence hypothesis. Additional passes through the alternative hypotheses are acceptable, provided the data are compared with comparable data for the top-ranked hypothesis.

System response timing statistics should be reported.

Data resulting from these tests is to be provided to NBS following the March meeting for detailed analysis and in evaluating alternative scoring software.

## DOCUMENTATION

Documentation on the characteristics of the imposed grammar(s) must be provided. This information should describe any use of the material from which the test material was drawn (i.e. the set of 2200 task domain sentences developed at BBN and used by TI in recording the Resource Management Speech Database).

The system architecture and hardware configuration used for these tests should be documented.

## REFERENCES

[1] D.S. Pallett, "Selected Test Material for the March 1987 DARPA Benchmark Tests",

Proceedings of the March 1987 DARPA Speech Recognition Workshop.

[2] (anonymous) "Integration, Transition and Performance Evaluation of Generic Artificial Intelligence Technology", Strategic Computing Program draft document dated Dec.6, 1985 (For Official Use Only).

[3] D.S. Pallett, "Benchmark Test Procedures for Continuous Speech Recognition Systems", draft document dated August 29, 1986 distributed prior to the Fall 1986 DARPA meeting.

[4] D.S. Pallett, "Performance Assessment of Automatic Speech Recognizers", Journal of Research of the National Bureau of Standards, Volume 90, Number 5, September-October 1985, pp. 371-387.

[5] A.I. Rudnicky, "Rules for Creating Lexicon Entries", note dated 11 February, 1987.